
Learning from Collective Intelligence in Groups

Guo-Jun Qi

QI4@ILLINOIS.EDU

Beckman Institute, University of Illinois at Urbana-Champaign, 405 N. Mathews, Urbana, IL 61801

Charu Aggarwal

CHARU@US.IBM.COM

IBM T.J. Watson Research Center, 1101 Kitchawan Road, Route 134, Yorktown Heights, NY 10598

Pierre Moulin

MOULIN@IFP.UIUC.EDU

Beckman Institute, University of Illinois at Urbana-Champaign, 405 N. Mathews, Urbana, IL 61801

Thomas Huang

HUANG@IFP.UIUC.EDU

Beckman Institute, University of Illinois at Urbana-Champaign, 405 N. Mathews, Urbana, IL 61801

Abstract

Collective intelligence, which aggregates the shared information from large crowds, is often negatively impacted by unreliable information sources with the low quality data. This becomes a barrier to the effective use of collective intelligence in a variety of applications. In order to address this issue, we propose a probabilistic model to jointly assess the reliability of sources and find the true data. We observe that different sources are often not independent of each other. Instead, sources are prone to be mutually influenced, which makes them dependent when sharing information with each other. High dependency between sources makes collective intelligence vulnerable to the overuse of redundant (and possibly incorrect) information from the dependent sources. Thus, we reveal the latent group structure among dependent sources, and aggregate the information at the group level rather than from individual sources directly. This can prevent the collective intelligence from being inappropriately dominated by dependent sources. We will also explicitly reveal the reliability of groups, and minimize the negative impacts of unreliable groups. Experimental results on real-world data sets show the effectiveness of the proposed approach with respect to existing algorithms.

1. Introduction

Collective intelligence aggregates contributions from multiple sources in order to collect data for a variety of tasks. For example, voluntary participants collaborate with each other to create a fairly extensive set of entries in *Wikipedia*; a crowd of paid persons may perform image and news article annotations in *Amazon Mechanical Turk*. These crowdsourced tasks usually involve multiple *objects*, such as Wikipedia entries and images to be annotated. The participating sources collaborate to claim their own *observations*, such as facts and labels, on these objects. Our goal is to aggregate these collective observations to infer the *true values* (e.g., the true fact and image label) for the different objects (Zhao et al., 2012; Pasternack & Roth, 2010; Galland et al., 2010).

We note that an important property of collective intelligence is that different sources are typically not independent of one another. For example, in the same social community, people often influence each other, where their judgments and opinions are not independent. In addition, task participants may obtain their data and knowledge from the same external information source, and their contributed information will be dependent. Thus, it may not be advisable to treat sources independently and directly aggregate the information from individual sources, when the aggregation process is clearly impacted by such dependencies. In this paper, we will infer the source dependency by revealing latent group structures among involved sources. Dependent sources will be grouped, and their reliability is analyzed at the group level. The incorporation of such dependency analysis in group structures can reduce the risk of overusing the observations made

by the dependent sources in the same group, especially when these observations are unreliable. This helps prevent dependent sources from inappropriately dominating collective intelligence especially when these source are not reliable.

Moreover, we note that groups are not equally reliable, and they may provide incorrect observations which conflict with each other, either unintentionally or maliciously. Thus, it is important to reveal the reliability of each group, and minimize the negative impact of the unreliable groups. For this purpose, we study the *general* reliability of each group, as well as its *specific* reliability on each individual object. These two types of reliability are closely related. General reliability measures the overall performance of a group by aggregating each individual reliability over the entire set of objects. On the other hand, although each object-specific reliability is distinct, it can be better estimated with a prior that a *generally reliable* group is likely to be reliable on an individual object and vice versa. Such prior can reduce the overfitting risk of estimating each object-specific reliability, especially considering that we need to determine the true value of each object at the same time (Kasneci et al., 2011; Bachrach et al., 2012).

The remainder of this paper is organized as follows. In Section 2, we formally define our problem and notations in the paper. The Multi-Source Sensing (MSS) model for the problem is developed in Section 3, followed by the group observation models in Section 4. Section 5 presents the inference algorithm. Then we evaluate the approach in Section 6 on real data sets, and conclude the paper in Section 7.

2. Problem and Notational Definitions

We formally define the following Multi-Source Sensing (MSS) model which abstracts the description of collective intelligence. Suppose that we have a set $\mathcal{S} := \{S_1, S_2, \dots, S_N\}$ of N sources, and a set $\mathcal{O} := \{O_1, O_2, \dots, O_M\}$ of M objects. Each object O_m takes a value t_m from a domain \mathcal{X}_m which describes one of its attributes. Each source S_n in \mathcal{S} reports its observation $y_{n,m} \in \mathcal{X}_m$ on an object O_m . Then the goal of the MSS model is to infer the true value t_m of each object O_m from the observations made by sources.

In this paper, we are particularly interested in categorical domain $\mathcal{X}_m = \{1, \dots, K_m\}$ with discrete values. For example, in many crowdsourcing applications, we focus on the (binary-valued) assertion correctness in hypothesis test and (multi-valued) categories in clas-

sification problem. However, the MSS model can be straightly extended to continuous domain. Due to the space limitation, we leave this topic in the extended version of this paper.

Figure 1 illustrates an example, where five sources make their observations on four objects. An object can be an image or a biological molecule, and an annotator or a biochemical expert (as a source) may claim the category (as the value) for each object. Alternatively, an object can be a book, and a book seller web site (as a source) claims the identity of its authors (as the values). In a broader sense, objects are even not concrete objects. They can refer to any crowdsourced tasks, such as questions (e.g., “is Peter a musician?”) and assertions (e.g., “George Washington was born on February 22, 1732.” and “an animal is present in an image,”), and the observations by sources are the answers to the questions, or binary-valued positive or negative claims on these assertions.

It is worth noting that each source does not need to claim the observations on all objects in \mathcal{O} . In many tasks, sources make claims only on small subsets of objects of interest. Thus, for notational convenience, we denote all claimed observations by \mathbf{y} in bold, and use $I = \{(n, m) | \exists y_{n,m} \in \mathbf{y}\}$ to denote all the indices in \mathbf{y} . We use the notations $I_{n,\cdot} = \{m | \exists (n, m) \in I\}$ and $I_{\cdot,m} = \{n | \exists (n, m) \in I\}$ to denote the subset of indices that are consistent with the corresponding subscripts n and m .

Meanwhile, to model the dependency among sources, we assume that there are a set of latent groups $\{G_1, G_2, \dots\}$, and each source S_n is assigned to one group G_{g_n} where $g_n \in \{1, 2, \dots\}$ is a random variable indicating its membership. For example, as illustrated in Figure 1, the five sources are inherently drawn from two latent groups, where each source is linked to the corresponding group by dotted lines. Each latent group contains a set of sources which are influenced by each other and tend to make similar observations on objects. The unseen variables of group membership will be inferred mathematically from the underlying observations. Here, we do not assume any prior knowledge on the number of groups. The composition of these latent groups will be determined with the use of a Bayesian nonparametric approach by stick-breaking construction (Sethuraman, 1994), as to be presented in the next section.

To minimize the negative impact of unreliable groups, we will explicitly model the group-level reliability. Specifically, for each group G_l , we define a group reliability score $u_l \in [0, 1]$ in unit interval. This value measures the general reliability of the group over the

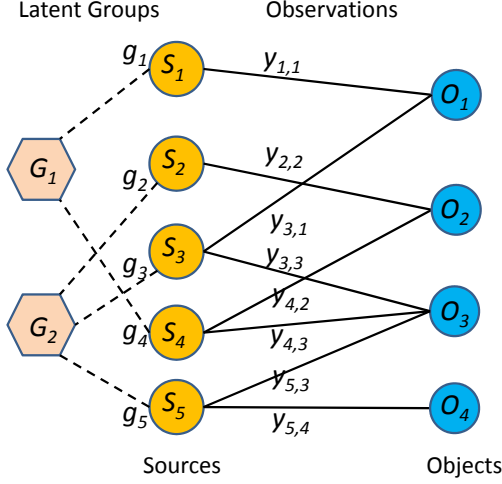


Figure 1. An example illustrating a set of five sources with their observations on four objects.

entire set of objects. The higher value of u_l indicates the greater reliability of the group. Meanwhile, we also specify the reliability $r_{l,m} \in \{0, 1\}$ of each group G_l on each particular object O_m . When $r_{l,m} = 1$, group G_l will have reliable performance on O_m , and otherwise it will be unreliable. In the next section, we will clarify the relationship between general reliability u_l and object-specific reliability $r_{l,m}$.

3. Multi-Source Sensing Model

In this section, we present a generative process for the multi-source sensing problem. It defines a group reliability structure to find the dependency between sources at the same time when we infer their reliability at the group level.

First we define the following generative model for multi-source sensing (MSS) process below, the details of which will be explained shortly.

$$\boldsymbol{\lambda} \sim \text{GEM}(\kappa), \quad g_n | \boldsymbol{\lambda} \sim \text{Discrete}(\boldsymbol{\lambda}), \quad (1)$$

$$u_l \sim \text{Beta}(b_1, b_0), r_{l,m} \sim \text{Bern}(u_l), t_m \sim \text{Unif} \quad (2)$$

$$\boldsymbol{\pi}_{l,m} | r_{l,m}, t_m = z \sim H_{r_{l,m}}(t_m) \quad (3)$$

$$y_{n,m} | \boldsymbol{\pi}_{l,m}, g_n \sim F(\boldsymbol{\pi}_{g_n,m}) \quad (4)$$

for $n = 1, 2, \dots, N, m = 1, 2, \dots, M, l = 1, 2, \dots$. Figure 2 illustrates the generative process in a graphical representation. Here, $g_n | \boldsymbol{\lambda} \sim \text{Discrete}(\boldsymbol{\lambda})$ denotes a discrete distribution, which generates the value $g_n = i$ with probability λ_i ; Beta, Bern and Unif stand for Beta, Bernoulli and uniform distributions, respectively. We explain the detail of this generative process below.

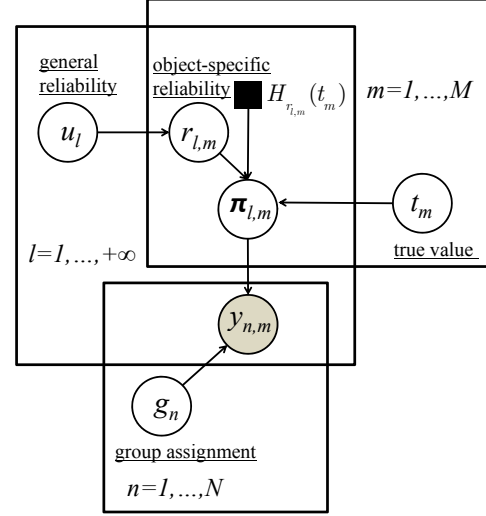


Figure 2. The graphical model for multi-source sensing.

In Eq. (1), we adopt the stick-breaking construction $\text{GEM}(\kappa)$ (named after Griffiths, Engen and McCloskey) with concentration parameter $\kappa \in \mathbb{R}^+$ to define the prior distribution of assigning each source S_n to a latent group G_{g_n} (Sethuraman, 1994). Specifically, in $\text{GEM}(\kappa)$, a set of random variables $\boldsymbol{\rho} = \{\rho_1, \rho_2, \dots\}$ are independently drawn from the Beta distribution $\rho_i \sim \text{Beta}(1, \kappa)$. They define the mixing weights $\boldsymbol{\lambda}$ of the group membership component such that $p(g_n = l | \boldsymbol{\rho}) = \lambda_l = \rho_l \prod_{i=1}^{l-1} (1 - \rho_i)$. Obviously, by the above stick-breaking process, we do not need the prior knowledge of the number of groups. This number will be determined by capturing the degree of dependency between sources.

Clearly, we can see that the parameter κ in the above GEM construction plays the vital role of determining *a priori* the degree of dependency between sources. Actually, according to the GEM construction, we can verify that the probability of two sources S_n and S_m being assigned to the same group is

$$\begin{aligned} P(g_n = g_m) &= \sum_{l=1}^{+\infty} \mathbb{E}_{\boldsymbol{\lambda}} P(g_n = l | \boldsymbol{\lambda}) P(g_m = l | \boldsymbol{\lambda}) \\ &= \sum_{l=1}^{+\infty} \mathbb{E}_{\lambda_l} \lambda_l^2 = \sum_{l=1}^{+\infty} \mathbb{E}_{\rho_l} \rho_l^2 \prod_{i=1}^{l-1} \mathbb{E}_{\rho_i} (1 - \rho_i)^2 \\ &= \sum_{l=1}^{+\infty} \frac{2}{(1 + \kappa)(2 + \kappa)} \left(\frac{\kappa}{2 + \kappa} \right)^{l-1} = \frac{1}{1 + \kappa} \end{aligned} \quad (5)$$

We can find that when κ is smaller, source are more likely to be assigned to the same group where they are dependent and share the same observation model. This will yield higher degree of dependency between sources. As κ increases, the probability that any two

sources belong to the same group will decrease. In the extreme case, as $\kappa \rightarrow +\infty$, this probability will approach to zero. In this case, all sources will be assigned to distinctive groups, yielding complete independence between sources. This shows that the model can flexibly capture the various degree of dependency between sources by setting an appropriate value of κ .

In Eq. (2), we define a Beta distribution $\text{Beta}(b_1, b_0)$ on the group reliability score u_l , where b_1 and b_0 are the soft counts which specify whether a group is reliable or not a priori, respectively. Then object-specific reliability $r_{l,m} \in \{0, 1\}$ is sampled from the Bernoulli distribution $\text{Bern}(u_l)$ to specify the group reliability on a particular object O_m . We can find that the higher the general reliability u_l , the more likely G_l is reliable on a particular object O_m with $r_{l,m}$ being sampled to be 1. This suggests that a generally more reliable group is more likely to be reliable on a particular object. In this sense, the general reliability serves as a prior to reduce the overfitting risk of estimating object-specific reliability in MSS model.

In Eq. (2), we adopt uniform distribution as the prior on the true value t_m of each object over its domain \mathcal{X}_m . The uniform distribution sets an unbiased prior so that true values will be completely determined a posteriori given observations in the model inference.

Eq. (3) and Eq. (4) define the generative process for the observations of each source in its assigned group. Specifically, given the group membership g_n , each source S_n generates its observation $y_{n,m}$ according to the corresponding group observation model $F(\pi_{g_n,m})$. The $\pi_{l,m}$ of this model is drawn from the conjugate prior $H_{r_{l,m}}(t_m)$ which depends on the true value t_m and the object-specific group reliability $r_{l,m}$. In the next section, we will detail the specification of $H_{r_{l,m}}(t_m)$ and $F(\pi_{l,m})$ in categorical domain. The models in other domain can be obtained by adopting the corresponding distribution with the analogous idea.

4. Group Observation Models

In categorical domain, for each group, we choose the multinomial distribution $F(\pi_{l,m}) = \text{Mult}(\pi_{l,m})$ as its observation model to generate observations $y_{n,m}$ for its member sources. Its parameter $\pi_{l,m}$ is generated by:

$$\pi_{l,m} | r_{l,m}, t_m = z \sim H_{r_{l,m}}(t_m) \\ := \text{Dir}(\underbrace{\theta^{(r_{l,m})}, \dots, \theta^{(r_{l,m})}}_{z-1}, \underbrace{\eta^{(r_{l,m})}}_{z^{\text{th entry}}, \dots, \theta^{(r_{l,m})}})$$

where Dir denotes Dirichlet distribution, and $\theta^{(r_{l,m})}$ and $\eta^{(r_{l,m})}$ are its soft counts for sampling the false and true values under different settings of $r_{l,m}$. Below we will explain how to set these soft counts under these settings.

For a reliable group G_l on object O_m (i.e., $r_{l,m} = 1$), it should be more likely to sample the true value $t_m = z$ as its observation than sampling any other false values. Thus, we should set a larger value for $\eta^{(r_{l,m})}$ than for $\theta^{(r_{l,m})}$.

On the other hand, if group G_l is unreliable on object O_m (i.e., $r_{l,m} = 0$), we can distinguish between *careless* and *malicious* groups, and set their parameters in different ways:

I. careless group: We define G_l as a careless group, whose member sources randomly claim values for object O_m , no matter which value is true. In this case, an equal soft count is set for the true and false values, i.e., $\theta^{(r_{l,m})} = \eta^{(r_{l,m})}$. This will make the true value indistinguishable from the false ones, so that the member sources makes a random guess of the true value.

II. malicious group: In this case, group G_l contains malicious sources which intentionally provide misleading information about the true value of object O_m . In other words, the group tends to claim the false values for object O_m , and thus we should set a larger value for $\theta^{(r_{l,m})}$ than for $\eta^{(r_{l,m})}$. Such malicious group can still contribute certain information if we read its observations in a reverse manner. Actually, by setting $\theta^{(r_{l,m})} > \eta^{(r_{l,m})}$, the MSS model gives the unclaimed observations larger weight (corresponding to larger value of $\theta^{(r_{l,m})}$) to be evaluated as the true value.

5. Model Inference

In this section, we present the inference and learning processes. The MSS model defines a joint distribution on $\mathbf{g} = \{g_n\}$, $\mathbf{r} = \{r_{l,m}\}$, $\mathbf{u} = \{u_l\}$, $\mathbf{t} = \{t_m\}$, $\boldsymbol{\pi} = \{\pi_{l,m}\}$ and the source observations \mathbf{y} . We wish to infer the tractable posterior $p(\mathbf{g}, \mathbf{r}, \mathbf{u}, \mathbf{t}, \boldsymbol{\pi} | \mathbf{y})$ with a parametric family of variational distributions in the factorized form:

$$q(\mathbf{g}, \mathbf{r}, \mathbf{u}, \mathbf{t}, \boldsymbol{\pi}) = \prod_n q(g_n | \varphi_n) \prod_{l,m} q(r_{l,m} | \tau_{l,m}) \\ \prod_l q(u_l | \beta_l) \prod_m q(t_m | \nu_m) \prod_{l,m} q(\pi_{l,m} | \alpha_{l,m})$$

with parameters φ_n , $\tau_{l,m}$, β_l , ν_m and $\alpha_{l,m}$ for these factors. The distribution and the parameter for each factor can be determined by variational approach (Jordan et al., 1999). Specifically, we aim to maximize the lower bound of the log likelihood $\log p(\mathbf{y})$,

i.e., $\mathcal{L}(q) = \mathbb{E}_q \ln p(\mathbf{g}, \mathbf{r}, \mathbf{u}, \mathbf{t}, \boldsymbol{\pi}, \mathbf{y}) - \mathbb{H}(q(\mathbf{g}, \mathbf{r}, \mathbf{u}, \mathbf{t}, \boldsymbol{\pi}))$ with the entropy function $\mathbb{H}(\cdot)$ to obtain the optimal factorized distribution. The lower bound can be maximized over one factor while the others are fixed. This is an approach which is similar to coordinate descent. All the factors are updated sequentially over steps until convergence. We derive the details of the steps for updating each factor below.

1: Update each factor $q(\boldsymbol{\pi}_{l,m}|\boldsymbol{\alpha}_{l,m})$ for the group observation parameter $\boldsymbol{\pi}_{l,m}$. By variational approach, we can verify that the optimal $q(\boldsymbol{\pi}_{l,m}|\boldsymbol{\alpha}_{l,m})$ has the form

$$\begin{aligned} q(\boldsymbol{\pi}_{l,m}|\boldsymbol{\alpha}_{l,m}) &\propto \exp\left\{\mathbb{E}_{q(r_{l,m}), q(t_m)} \ln p(\boldsymbol{\pi}_{l,m}|r_{l,m}, t_m)\right. \\ &\quad \left. + \sum_{n \in I_{l,m}} \mathbb{E}_{q(\mathbf{g}_n)} \ln p(y_{n,m}|\boldsymbol{\pi}_{l,m}, g_n)\right\} \\ &\propto \prod_{k \in \mathcal{X}} \pi_{l,m;k}^{\alpha_{l,m;k}-1} \end{aligned}$$

It still has Dirichlet distribution with the parameters

$$\begin{aligned} \alpha_{l,m;k} &= \sum_{n \in I_{l,m}} q(g_n = l) \delta[y_{n,m} = k] \\ &\quad + \sum_{r_{l,m} \in \{0,1\}} q(r_{l,m}) [(\eta^{(r_{l,m})} - 1) q(t_m = k) \\ &\quad + (\theta^{(r_{l,m})} - 1)(1 - q(t_m = k))] + 1 \end{aligned}$$

for each $k \in \mathcal{X}_m$, where $\delta[A]$ is the indicator function which outputs 1 if A holds, and 0 otherwise. Here we index the element in $\boldsymbol{\alpha}_{l,m}$ and $\boldsymbol{\pi}_{l,m}$ by k after the colon. We will follow this notation convention to index the element in vectors in this paper.

2: Update each factor $q(u_l|\beta_l)$ for general group reliability u_l . We have

$$\begin{aligned} \ln q(u_l|\beta_l) &\propto \sum_m \mathbb{E}_{q(r_{l,m})} \ln p(r_{l,m}|u_l) + \ln p(u_l|b_1, b_0) \\ &= \left(\sum_m q_1(r_{l,m}) + b_1 - 1\right) \ln u_l \\ &\quad + \left(\sum_m q_0(r_{l,m}) + b_0 - 1\right) \ln(1 - u_l) \end{aligned}$$

where $q_i(r_{l,m})$ is short for $q(r_{l,m} = i)$ for $i = 0, 1$, respectively. We can find the posterior of u_l still has Beta distribution as $\text{Beta}(\beta_l)$ with parameter

$$\beta_l = \left[\sum_m q_1(r_{l,m}) + b_1, \sum_m q_0(r_{l,m}) + b_0\right].$$

We can find that the above updated parameter sums up the posterior reliability $q_1(r_{l,m})$ and $q_0(r_{l,m})$ over all objects. This corresponds with the intuition that

the general reliability is the sum of the reliability on individual objects.

3: Update each factor $q(r_{l,m}|\tau_{l,m})$ for the object-specific reliability $r_{l,m}$ of group G_l on O_m :

$$\begin{aligned} \ln q(r_{l,m}|\tau_{l,m}) &\propto \mathbb{E}_{q(t_m), q(\boldsymbol{\pi}_{l,m})} \ln q(\boldsymbol{\pi}_{l,m}|r_{l,m}, t_m) \\ &\quad + \mathbb{E}_{q(u_l)} \ln q(r_{l,m}|u_l) \end{aligned} \quad (6)$$

Thus, we have

$$\begin{aligned} &\ln q(r_{l,m}|\tau_{l,m}) \\ &\propto \sum_{k \in \mathcal{X}_m} q(t_m = k) [(\eta^{(r_{l,m})} - 1) \mathbb{E}_{q(\boldsymbol{\pi}_{l,m})} \ln \pi_{l,m;k} \\ &\quad + (\theta^{(r_{l,m})} - 1) \sum_{j \neq k} \mathbb{E}_{q(\boldsymbol{\pi}_{l,m})} \ln \pi_{l,m;j}] \\ &\quad + r_{l,m} \mathbb{E}_{q(u_l)} \ln u_l + (1 - r_{l,m}) \mathbb{E}_{q(u_l)} \ln(1 - u_l) \end{aligned} \quad (7)$$

for $r_{l,m} \in \{0, 1\}$, respectively. Here we compute the expectation of the logarithmic Dirichlet variable as

$$\mathbb{E}_{q(\boldsymbol{\pi}_{l,m})} \ln \pi_{l,m;k} = \psi\left(\sum_i \alpha_{l,m;i}\right) - \psi(\alpha_{l,m;k})$$

with the digamma function $\psi(\cdot)$; the expectation of the logarithmic Beta variables

$$\mathbb{E}_{q(u_l)} \ln u_l = \psi(\beta_{l;1} + \beta_{l;2}) - \psi(\beta_{l;1})$$

and

$$\mathbb{E}_{q(u_l)} \ln(1 - u_l) = \psi(\beta_{l;1} + \beta_{l;2}) - \psi(\beta_{l;2}).$$

Finally, the updated values of $q(r_{l,m})$ are normalized to be valid probabilities.

The last line of Eq. (7) reflects how the general reliability u_l affects the estimation of the object-specific reliability. This embodies the idea that a generally reliable group is likely to be reliable on a particular object and vice versa. This can reduce the overfitting risk of estimating $r_{l,m}$ especially considering that $q(t_m)$ in the second line also need to be estimated simultaneously in MSS model as in the next step.

4: Update each factor $q(t_m|\nu_m)$ for the true value. We have

$$\begin{aligned} \ln q(t_m = k|\nu_m) &\propto \ln p(t_m = k) \\ &\quad + \sum_l \sum_{r_{l,m} \in \{0,1\}} q(r_{l,m}) \mathbb{E}_{q(\boldsymbol{\pi}_{l,m})} \ln p(\boldsymbol{\pi}_{l,m}|t_m = k, r_{l,m}) \end{aligned}$$

This suggests that

$$\begin{aligned} &\ln q(t_m = k|\nu_m) \\ &\propto \sum_l \sum_{r_{l,m}} q(r_{l,m}) \{(\eta^{(r_{l,m})} - 1) \mathbb{E}_{q(\boldsymbol{\pi}_{l,m})} \ln \pi_{l,m;k} \\ &\quad + \sum_{k' \neq k} (\theta^{(r_{l,m})} - 1) \mathbb{E}_{q(\boldsymbol{\pi}_{l,m})} \ln \pi_{l,m;k'}\} \end{aligned}$$

All $q(t_m = k), k \in \mathcal{X}_m$ are normalized to ensure they are validate probabilities.

5: Update each factor $q(g_n | \varphi_n)$ for the group assignment of each source. We can derive

$$\begin{aligned} & \ln q(g_n = l | \varphi_n) \\ & \propto \mathbb{E}_{q(\rho)} \ln p(g_n = l | \rho) + \sum_{m \in I_n, \cdot} \mathbb{E}_{q(\pi_{l,m})} \ln p(y_{n,m} | \pi_{l,m}, g_n = l) \\ & = \mathbb{E}_{q(\rho)} \ln p(g_n = l | \rho) + \sum_{m \in I_n, \cdot} \mathbb{E}_{q(\pi_{l,m})} \ln \pi_{l,m; y_{n,m}} \end{aligned}$$

This shows that $q(g_n = l | \varphi_n)$ is a multinomial distribution with its parameter as

$$\varphi_{n;l} = q(g_n = l | \varphi_n) = \frac{\exp(U_{n,l})}{\sum_{l=1}^{\infty} \exp(U_{n,l})} \quad (8)$$

where

$$U_{n,l} = \mathbb{E}_{q(\rho)} \ln p(g_n = l | \rho) + \sum_{m \in I_n, \cdot} \mathbb{E}_{q(\pi_{l,m})} \ln \pi_{l,m; y_{n,m}}$$

As in (Kurihara et al., 2006), we truncate after L groups: the posterior distribution $q(\rho_i)$ after the level L is set to be its prior $p(\rho_i)$ from Beta(1, κ); and all the expectations $\mathbb{E}_{q(\pi_{l,m})} \ln \pi_{l,m;k}$ after L are set to:

$$\mathbb{E}_{q(\pi_{l,m})} \ln \pi_{l,m;k} = \mathbb{E}_{q(t_m), p(r_{l,m})} \{ \mathbb{E}[\ln \pi_{l,m;k} | r_{l,m}, t_m] \}$$

with $p(r_{l,m})$ defined as in Eq. (2) for all $l > L$, respectively. The inner conditional expectation in the above is taken with respect to the probability of $\pi_{l,m}$ conditional on $r_{l,m}$ and t_m as defined in (3). Similar to the family of nested Dirichlet process mixture in (Kurihara et al., 2006), this will form a family of nested priors indexed by L for the MSS model. Thus, we can compute the infinite sum in the denominator of Eq. (8) as:

$$\sum_{l=L+1}^{\infty} \exp(U_{n,l}) = \frac{\exp(U_{n,L+1})}{1 - \exp(\mathbb{E}_{\rho_i \sim \text{Beta}(1, \kappa)} \ln(1 - \rho_i))}$$

6: Finally, we can find that before the truncation level L , the posterior distribution $q(\rho_i) \sim \text{Beta}(\phi_{i,1}, \phi_{i,2})$ is updated as

$$\phi_{i,1} = 1 + \sum_{n=1}^N q(g_n = i), \quad \phi_{i,2} = \kappa + \sum_{n=1}^N \sum_{j=i+1}^{\infty} q(g_n = j)$$

The above steps are iterated to yield the optimal factors.

6. Experimental Results

In this section, we compare our approach with other existing algorithms and demonstrate its effectiveness for inferring source reliability together with the true values of objects. The comparison is performed on a book author data set from online book stores, and a user tagging data set from the online image sharing web site [Flickr.com](https://www.flickr.com/).

Book author data set: The first data set is the book author data set prepared in (Yin et al., 2007). The data set is obtained by crawling 1,263 computer science books on *AbeBooks.com*. For each book, *AbeBooks.com* returns the book information extracted from a set of online book stores. This data set contains a total of 877 book stores (sources), and 24,364 listings of books (objects) and their author lists (object values) reported by these book stores. Note that each book has a different categorical domain \mathcal{X} that contains all the authors claimed by sources.

Author names are normalized by preserving the first and last names, and ignoring the middle name of each author. For evaluation purposes, the authors of 100 books are manually collected from the scanned book covers (Yin et al., 2007). We compare the returned results of each model with the ground truth author lists on this test set and report the accuracy.

We compare the proposed algorithm with the following ones: (1) the naive Voting algorithm which counts the top voted author list for each book as the truth; (2) *TruthFinder* (Yin et al., 2007); (3) *Accu* (Dong et al., 2009) which considers the dependency between sources; (4) *2-Estimates* as described in (Galland et al., 2010) with the highest accuracy among all the models in (Galland et al., 2010) (5) *MSS*, which is our proposed algorithm. In the experiments, we choose the parameters $\eta^{(r_{l,m})}$ and $\theta^{(r_{l,m})}$ from $\{1.0, 2.0, 5.0, 10.0\}$ for $r_{l,m} \in \{0, 1\}$ as in Section 4, b_0, b_1 from $\{1.0, 2.0, 4.0\}$, and κ from $\{1.0, 5.0, 10.0\}$. Due to the unsupervised nature of the problem, we pick the set of parameters with the maximum observation likelihood.

Table 2 compares the results of the different algorithms on the book author data set in terms of the accuracy. The *MSS* model achieves the best accuracy among all the compared models. We note that the proposed *MSS* model is an unsupervised algorithm which does not involve any training data. That is to say, we do not use any true values in the *MSS* algorithm in order to produce the reliability ranking as well as other true values. Even compared with the accuracy of 0.91 of the Semi-Supervised Truth Finder (SSTF) (Yin & Tan, 2011)

Table 1. Top-10 and bottom-10 book stores ranked by their posterior probability of belonging to a reliable group. We also report the accuracy of these bookstores on the test set.

top-10 bookstore	accuracy	bottom-10 bookstore	accuracy
International Books	1	textbooksNow	0.0476
happybook	1	Gunter Koppon	0.225
eCampus.com	0.9375	www.textbooksrus.com	0.3333
COBU GmbH & Co. KG	0.875	Gunars Store	0.2308
HTBOOK	1	Indoo.com	0.3846
AlphaCraze.com	0.8462	Bobs Books	0.4615
Cobain LLC	1	OPOE-ABE Books	0
Book Lovers USA	0.8667	The Book Depository	0.3043
Versandantiquariat Robert A. Mueller	0.8158	Limelight Bookshop	0.3896
THESAINTBOOKSTORE	0.8214	textbookxdotcom	0.4444

Table 2. Comparison of different algorithms on book author and Flickr data set. On book author data set, the algorithms are compared by their accuracies. On Flickr data set, the algorithms are compared by their average precisions and recalls on 12 tags.

Model	book author data set	Flickr data set	
	accuracy	precision	recall
<i>Voting</i> (Dong et al., 2009)	0.71	0.8499	0.8511
<i>2-Estimates</i> (Galland et al., 2010)	0.73	0.8545	0.8602
<i>TruthFinder</i> (Yin & Tan, 2011)	0.83	0.8637	0.8649
<i>Accu</i> (Dong et al., 2009)	0.87	0.8731	0.8743
<i>MSS</i>	0.95	0.9176	0.9212

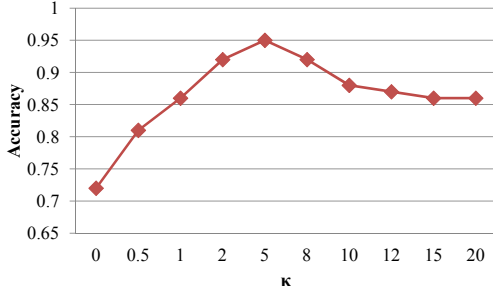


Figure 3. Model accuracy versus different κ on book author dataset.

using extra training data with known true values on some objects, the *MSS* model still achieves the highest 0.95 accuracy. It suggests that with additional training data, the *MSS* model may improve its accuracy further.

Since κ is predicative of the dependency between sources, we study the changes of the model accuracy versus various κ in Figure 3. We know that when $\kappa = 0$, all sources are completely dependent, and assigned to the same group. At this time, the model has a much lower accuracy, since all sources are tied to the same level of reliability within a single group. As κ

increases, the accuracy achieves the peak at $\kappa = 5.0$. After that, it deteriorates as the model gradually stops capturing the source dependency with increased κ . This demonstrates the importance of modeling the source dependency, and the capability of *MSS* model to capture such dependency by κ .

Moreover, to compare the reliability between sources, we can define the reliability of each source S_n by the expected reliability score of its assigned groups as

$$\text{Reliability}(S_n) = \sum_l q(g_n = l) \mathbb{E}_{q(u_l|\beta_l)}[u_l]$$

where

$$\mathbb{E}_{q(u_l|\beta_l)}[u_l] = \frac{\beta_{l,1}}{\beta_{l,1} + \beta_{l,2}}$$

Then, sources can be ranked based on such source reliability. In Table 1, we rank the top-10 and bottom-10 book stores in this way. In order to show the extent to which this ranking list is consistent with the real source reliability, we provide the accuracy of these bookstores on test data sets. Note that each individual bookstore may only claim on a subset of books in the test set, and the accuracy is computed based on the claimed books. From the table, we can see that the obtained rank of data sources is consistent with the rank of their accuracies on the test set. On the contrary, the accuracy

Table 3. The rounds used before convergence and computing time for each model.

Model	Bookstore		User Tagging	
	Rounds	Time(s)	Rounds	Time (s)
Voting	1	0.2	1	0.5
2-Estimates	29	21.2	32	628.1
TruthFinder	8	11.6	11	435.0
Accu	22	185.8	23	3339.7
MSS	9	10.3	12	366.2



Figure 4. Examples of image and the associated user tags in Flickr data set. In each subfigure the left image is correctly tagged by users, while the right one is wrongly tagged.

of the bottom-10 bookstores is much worse compared to that of the top-10 book stores on the test set. This also explains partly the better performance of the MSS model.

User tagging data set: We also evaluate the algorithm on a user tagging data set from an online image sharing web site *Flickr.com*. This data set contains 13,528 users (data sources) who annotate 36,280 images (data objects) with their own tags. We consider 12 tags - “balloon,” “bird,” “box,” “car,” “cat,” “child,” “dog,” “flower,” “snow leopard,” “waterfall,” “guitar,” “pumpkin” for evaluation purposes. Each tag is associated with a binary value 1/0 to represent its presence or not in an image, and we apply MSS model to these 12 tags separately to find whether they are present on each image. To test accuracy, we manually annotate these 12 tags on a subset of 1,816 images. Figure 4 illustrates some image examples in this data set and the tags annotated by users. We can find some images are wrongly tagged by users. The MSS model aims to correct these errors and yield accurate annotations on these images.

We follow the same experimental setup as on the book author data set. Table 2 shows the average precision and recall on the 12 tags by the compared algorithms. We can see that *MSS* still performs the best among these compared algorithms.

We also compare the computational time used by different algorithms in Table 3. The experiments are conducted on a personal computer with Intel Core i7-2600 3.40 GHz CPU, 8 GB physical memory and Windows 7 operating system. We can see that compared with most of other algorithms, MSS model can converge in fewer rounds with less computational cost.

7. Conclusion

In this paper, we propose an integrated true value inference and group reliability approach. Dependent sources which are grouped together, and their (general and specific) reliability is assessed at group level. The true data values are extracted from the reliable groups so that the risk of overusing the observations from dependent sources can be minimized. The overall approach is described by a probabilistic multi-source sensing model, based on which we jointly infer group reliability as well as the true values for objects *a posteriori* given the observations from sources. The key to the success of this model is to capture the dependency between sources, and aggregate the collective knowledge at the group granularity. We present experimental results on two real data sets, which demonstrate the effectiveness of the proposed model over other existing algorithms.

References

- Bachrach, Y., Minka, T., Guiver, J., and Graepel, T. How to grade a test without knowing the answers - a bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *Proc. of International Conference on Machine Learning*, 2012.
- Dong, X. L., Berti-Equille, L., and Srivastava, D. Integrating conflicting data: The role of source dependence. In *Proc. of International Conference on Very Large Databases*, August 2009.
- Galland, A., Abiteboul, S., Marian, A., and Senellart, P. Corroborating information from disagreeing

- views. In *Proc. of ACM International Conference on Web Search and Data Mining*, February 2010.
- Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.
- Kasneci, G., Gael, J. V., Stern, D., and Graepel, T. Cobayes: Bayesian knowledge corroboration with assessors of unknown areas of expertise. In *Proc. of ACM International Conference on Web Search and Data Mining*, 2011.
- Kurihara, K., Welling, M., and Vlassis, N. Accelerated variational dirichlet process mixtures. In *NIPS*, 2006.
- Pasternack, J. and Roth, D. Knowing what to believe (when you already know something). In *Proc. of International Conference on Computational Linguistics*, August 2010.
- Sethuraman, J. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.
- Yin, X. and Tan, W. Semi-supervised truth discovery. In *Proc. of International World Wide Web Conference*, March 28–April 1 2011.
- Yin, X., Han, J., and Yu, P. S. Truth discovery with multiple conflicting information providers on the web. In *Proc. of ACM SIGKDD conference on Knowledge Discovery and Data Mining*, August 2007.
- Zhao, B., Rubinstein, B. I. P., Gemmell, J., and Han, J. A bayesian approach to discovering truth from conflicting sources for data integration. In *Proc. of International Conference on Very Large Databases*, 2012.